FRONTERA, STAMPEDE, AND LONESTAR AT TACC

COMPUTE RESOURCES FOR THE GREATER SCIENTIFIC COMMUNITY

… AND THEN THERE IS ALSO HORIZON

Lars Koesterke
High Performance Computing
Texas Advanced Computing Center, TACC
The University of Texas at Austin

TAMU, March 18-19, 2023

- What is TACC?
- What is Frontera, Stampede2, and Lonestar6?
- What is Horizon?

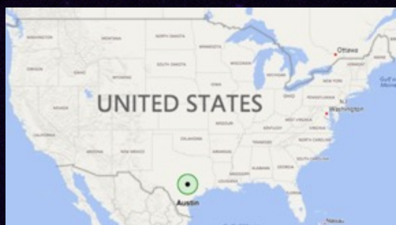
WHAT IS TACC?

Grendel, 1993

Frontera, 2019

The Texas Advanced Computing Center, at UT Austin is a (primarily) NSF-funded center to provide and apply large scale computing resources to the open science community.

TACC AT A GLANCE - 2022



Personnel

190 Staff (~70 PhD)

Facilities

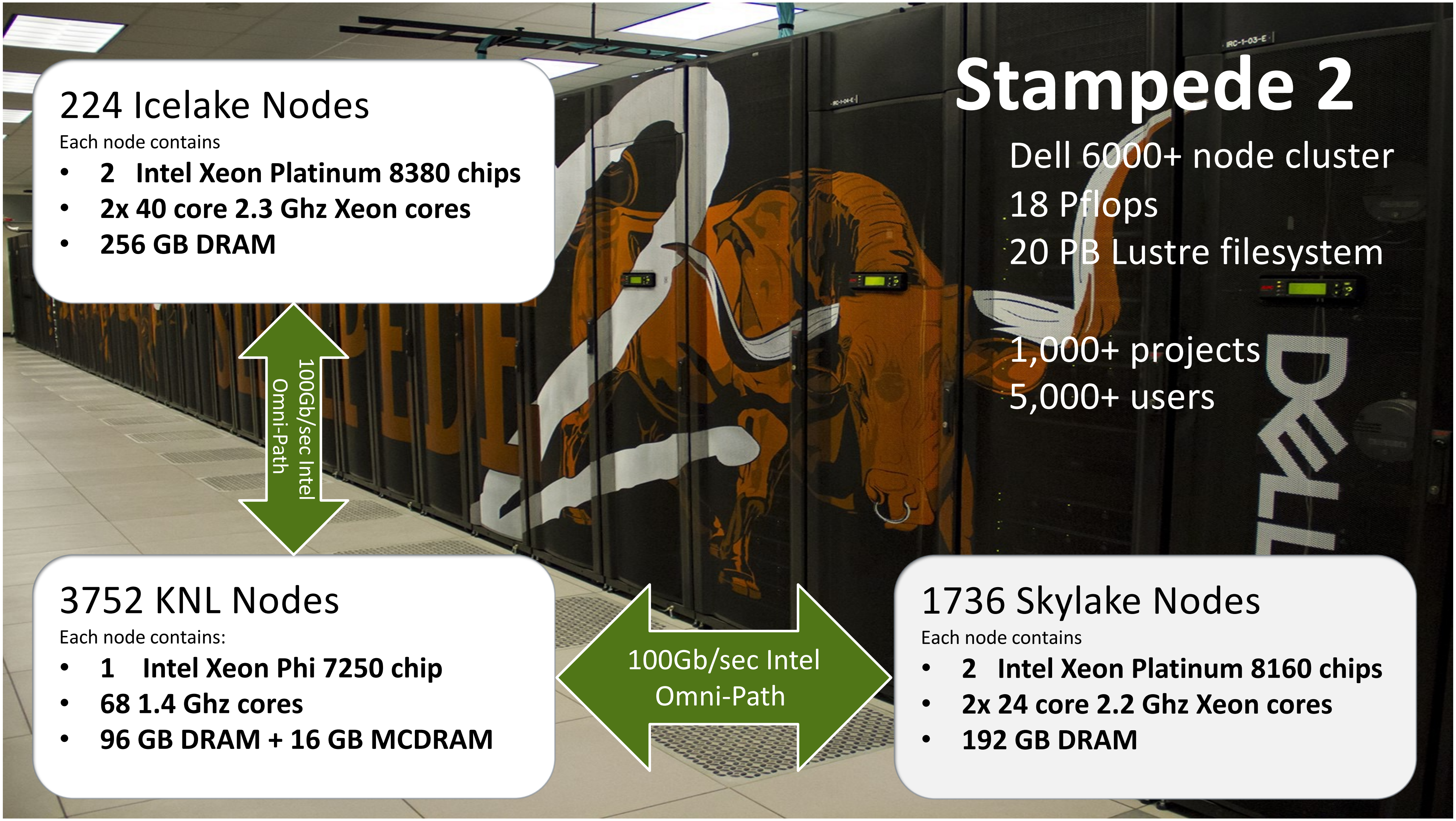
12 MW Data center capacity
Two office buildings, Three Datacenters, two visualization facilities, and a chilling plant.

Systems and Services

>Seven Billion compute hours per year
>5 Billion files, >100 Petabytes of Data,
NSF Frontera (Track 1), Stampede2 (ACCESS Flagship), Jetstream2 (Cloud), Chameleon (Cloud Testbed) system

Usage

>10,000 direct users in >1,000 projects, >50,000 web/portal users, User demand 8x available system time.
Thousands of training/outreach participants annually



Stampede 2

Dell 6000+ node cluster
18 Pflops
20 PB Lustre filesystem

1,000+ projects
5,000+ users

224 Icelake Nodes

Each node contains

- **2 Intel Xeon Platinum 8380 chips**
- **2x 40 core 2.3 Ghz Xeon cores**
- **256 GB DRAM**

100Gb/sec Intel Omni-Path

3752 KNL Nodes

Each node contains:

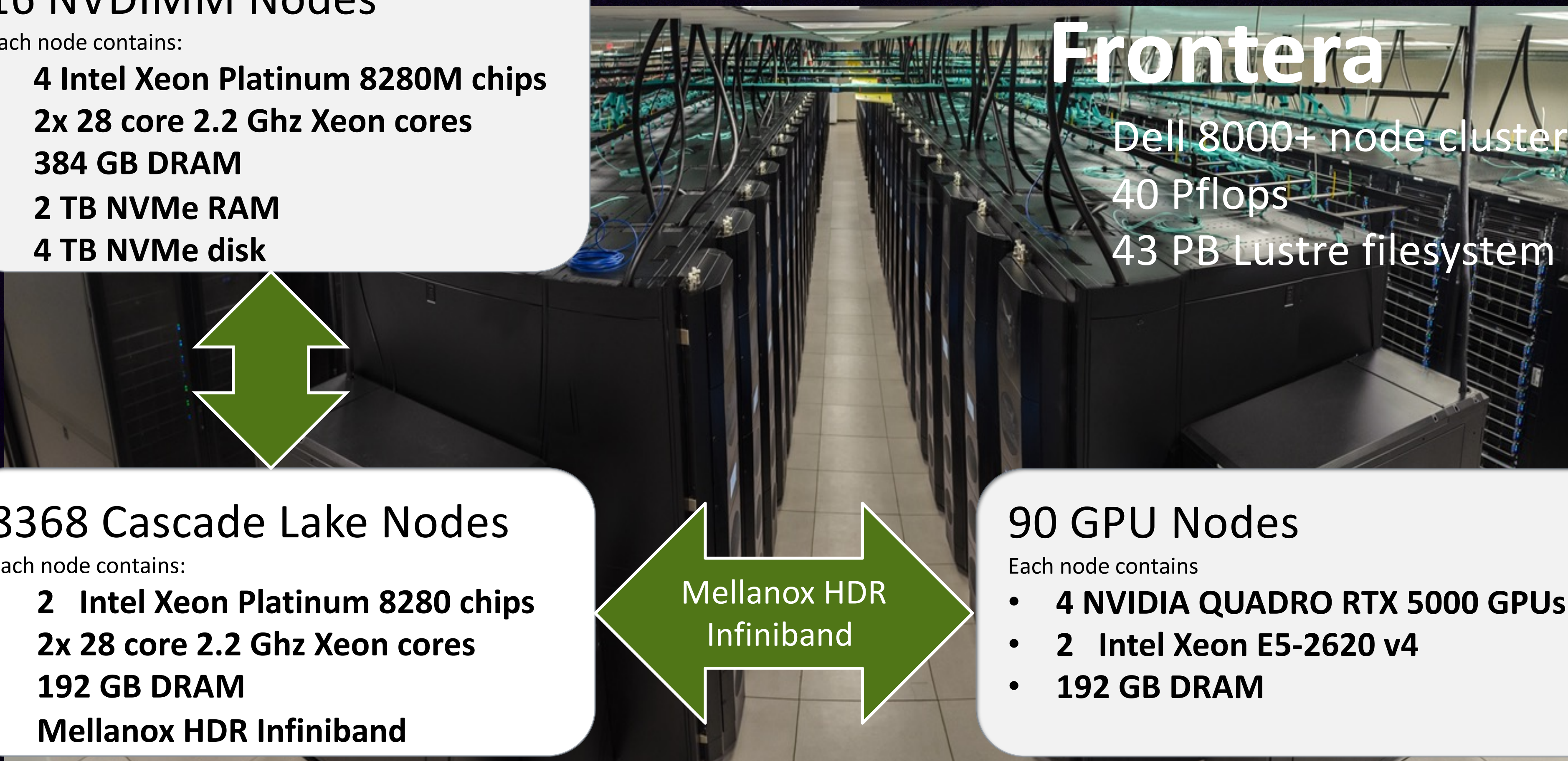
- **1 Intel Xeon Phi 7250 chip**
- **68 1.4 Ghz cores**
- **96 GB DRAM + 16 GB MCDRAM**

100Gb/sec Intel Omni-Path

1736 Skylake Nodes

Each node contains

- **2 Intel Xeon Platinum 8160 chips**
- **2x 24 core 2.2 Ghz Xeon cores**
- **192 GB DRAM**



Frontera

Dell 8000+ node cluster
40 Pflops
43 PB Lustre filesystem

16 NVDIMM Nodes

Each node contains:

- **4 Intel Xeon Platinum 8280M chips**
- **2x 28 core 2.2 Ghz Xeon cores**
- **384 GB DRAM**
- **2 TB NVMe RAM**
- **4 TB NVMe disk**

8368 Cascade Lake Nodes

Each node contains:

- **2 Intel Xeon Platinum 8280 chips**
- **2x 28 core 2.2 Ghz Xeon cores**
- **192 GB DRAM**
- **Mellanox HDR Infiniband**

Mellanox HDR Infiniband

90 GPU Nodes

Each node contains

- **4 NVIDIA QUADRO RTX 5000 GPUs**
- **2 Intel Xeon E5-2620 v4**
- **192 GB DRAM**

THIRD YEAR OF PRODUCTION ON FRONTERA

- In the last 12 months:
  - Uptime of 99.2%
  - Average Utilization of 95.4%
  - ~72M SUs delivered
  - 1.13M jobs delivered
  - Zero security incidents.
- On the bright side, we are always full.
- On the downside, no way to squeeze anything else in.

USAGE ON FRONTERA

- **>2,000 jobs were >25,000 cores** – about a **quarter** of all cycles on large jobs.
- >100 jobs at half or full system scale (Consider if all jobs were full scale, and averages 24 hours, we'd only run 365 jobs a year, as opposed to 1.1M jobs).
- Flex jobs, used for backfill, represent 20% of the jobs run (>200k), but represent less than 0.5% of SUs delivered (285k out of 70M).
- Small jobs represent ~30% of jobs, but less than 2% of cycles delivered.
  - So **97% of time goes to jobs >2 nodes**.
  - Average jobs size about **6x that of Stampede2** – this machine *is* used differently.
- We tune the scheduling policy multiple times a year… essentially adjusting to demand.

TEXASCALE DAYS

- Opportunity to run at full or half scale
  - 8k nodes – full
  - 4k nodes – half
  - Dedicated access
- Production
  - 24 hours
- Benchmark
  - 2 hour blocks

ALLOCATIONS

- Three Main Tracks:
  - **Leadership Resource Allocations** – Ready to run at large scale, 250k-5M node-hours per year. (Currently 49 projects active)
  - **Pathways Allocations –** Not yet at that scale, but scientific potential to get there, up to 200k node hours.
  - **Large Scale Community Partnerships** - For Gateways, Community Codes, or large scientific collaborations, up to 3 years, 25k-1M hours per year.
- Also Startup, Educational, and Discretionary allocations

ALLOCATIONS

	Leadership Resource Allocation (LRAC)	Pathways	Large-scale Community Partnerships (LSCP)	Totals
Requests	62	38	13	113
Unique Pis	62	37	12	103
Unique Orgs	45	31	12	69
CPU SUs Requested	93,914,427	5,634,333	12,377,780	111,926,540
CPU SUs Awarded	54,076,692	3,433,463	6,086,100	63,596,255
GPU SUs Requested	2,545,684	439,665	1,385,600	4,370,949
GPU SUs Awarded	613,684	146,975	160,000	920,659

HOW MUCH DOES IT COST ME TO ACCESS?
(YOU WILL LIKE THE ANSWER)

- TACC’s HPC resources and staff expertise are available to the research community usually at no additional cost. We are funded mostly by the US National Science Foundation to support open academic research.
- International collaborations are welcome.
- Access is free

LONESTAR

- **Lonestar6** – Research computing for Univ of Texas System and other paying partners.
- **Texas A&M** is one of the partners
  - Also participation in Lonestar5 and Lonestar4
- A&M Allocation managed by TAMU

- Lonestar6
  - AMD Dual-socket
  - Nodes immersed in oil

FRONTERA/LONESTAR

- Frontera: 8,360 primary compute nodes – 40PF, >1.5PB of RAM, 60PB scratch, 3PB fast (flash) scratch, fast interconnect.
  - 2 Intel Cascade Lake processors, 56 cores, 192GB of RAM per node.
  - Normal production runs to 2k nodes which is >100k cores.
  - "Texascale" runs to 8k nodes/450k cores.
  - 16 NVDIMM nodes – 6TB of RAM or fast storage.
  - 90 4x GPU nodes – 360 RTX 5000 oil-cooled GPUs.
- **Lonestar6**
  - **560 nodes, each with 2 AMD EPYC 64-core processors.**
  - **GPU subsystem – 80 nodes, 3 NVIDIA Ampere A100 GPUs per node (120GB GPU memory, 256GB main memory per node).**

AND THEN THERE IS HORIZON

LEADERSHIP CLASS COMPUTING FACILITY (LCCF)

- In 2018 when Frontera was awarded by the NSF, this also included the opportunity to develop the proposal for the LCCF. The first HPC system in the ***LCCF must be 10x more capable than Frontera***.
- The LCCF will be awarded in the operations side of the NSF, the Large Facilities Office. This is the first time that NSF will be funding HPC as part of operations.
- We will be submitting the third and final proposal of the 3-phase process that takes multiple years to complete.
- Pending successful review, and subsequent funding action, we anticipate starting the 2-year construction phase in 2024.

THE 15MW FACILITY WILL HOST HORIZON

Artist Rendering of Switch Data Centers at Dell Headquarters

HORIZON, OUR NEXT MACHINE (SO WE HOPE)

- What should the machine look like?
  - Nothing really exotic like a Quantum Computer
- **Triangle**
  - Usability
  - Costs: $ per Flop, and $ per ‘power unit’
  - Software environment
  - … and then there is Machine Learning as well

USABILITY (1)

- We do not pretend to have all the answers right away
- We use actual data (as much as possible)
- We talk to:
  - Users
  - Stakeholders
  - Other centers
  - Our own staff
  - Characteristic Science Applications (CSA), more on this later

USABILITY (2)

- For users it is often a zero sum game
  - *'If I can use the machine well that is good'*
  - *'If others cannot use the machine … even better'*

- For us it is not a zero-sum game
- Maximize scientific output
- Happy users

COSTS

- GPUs are better in some metrics
  - $ per Flop: of the order of 2x
  - $ per 'power unit': of the order of 3x
- Apples to Oranges (for scientific applications, not ML)
  - Compare 1 GPU to 1 dual-socket CPU node
  - Roughly the same are
  - Number of transistors: GPU has more
  - Power consumption: GPU a bit higher
  - Raw Flops (excluding tensor core): GPU significantly higher (3x)
  - Cost: GPU a bit more expensive
- Break even when: 1 GPU is about 1.5x faster than a dual-socket CPU node
  - Speed-ups of 50x for scientific applications are usually bogus
  - These numbers are often derived from 1 core vs. 1 GPU

SOFTWARE ENVIRONMENT

- Mature environment
- Compilers, debuggers, profilers, performance libraries, etc.
- CPU: MPI + X
  - Threads, most likely OpenMP
- GPU: MPI + Y, also threads, but there are many options
  - OpenMP
  - OpenACC
  - CUDA
  - HIP
  - SYCL with C++
  - (OpenCL)

WHAT DO OTHER PROVIDERS DO?

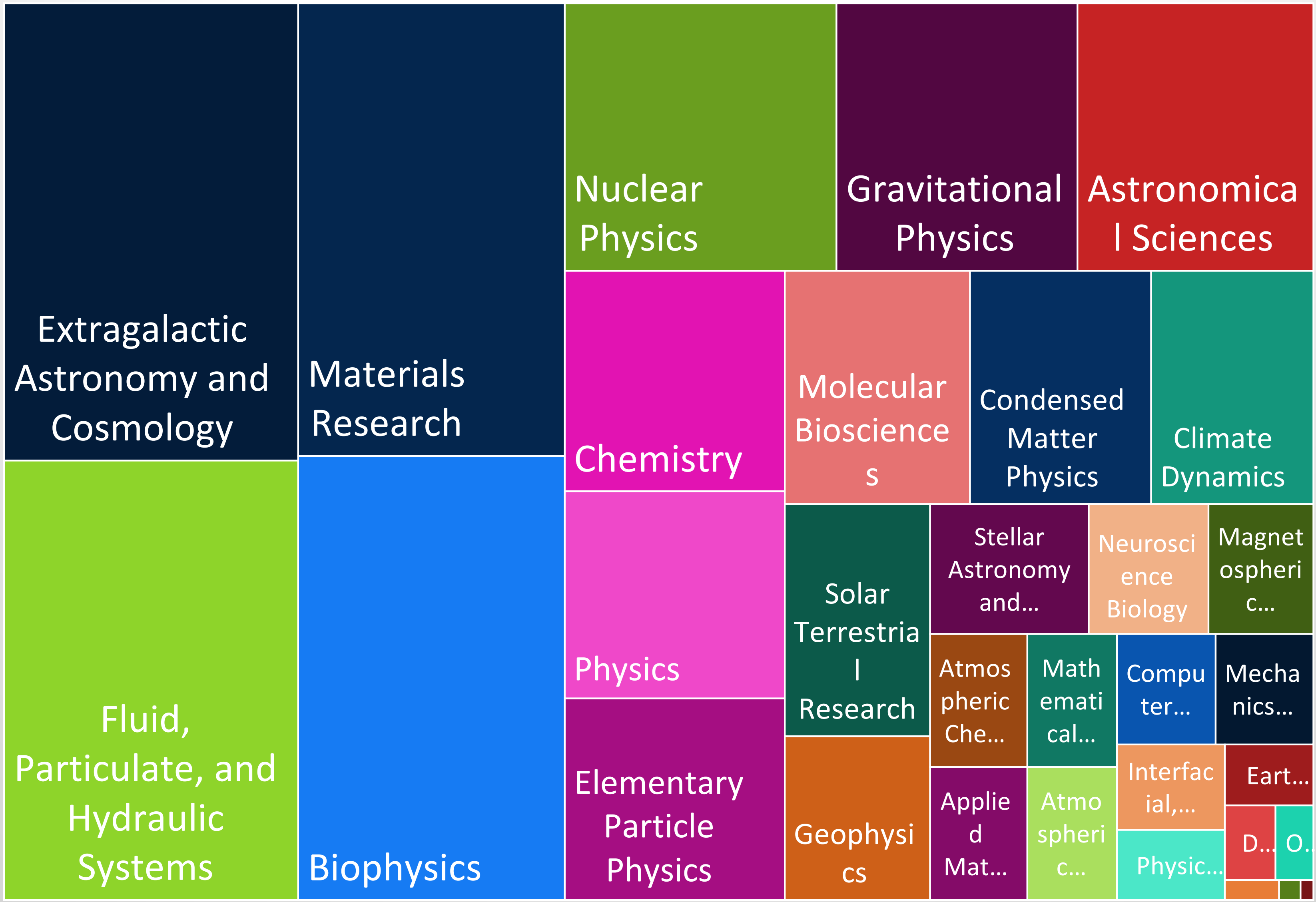
- DOE: The largest machines are 98-99% GPU
  - Few applications, so it seems
  - Lots of money for porting software and software development
- Would that work for us?

- We want to maximize scientific output
- What that really means is: We want happy users!
  - A large number across a broad spectrum of applications and fields of science
  - PIs won't (be able to) hire dedicated 'Software specialists' for code development

- We use Characteristic Science Applications (CSA)

FIELDS OF SCIENCE

- From last allocation request



TACC TEXAS

CHARACTERISTIC SCIENCE APPLICATIONS (CSA)

CSAs are initiated with the following three elements

- Application – science code or workflow
- Challenge problem – problem that cannot be readily solved today
- Figure of Merit (F.O.M.) – measure of performance of the application

The goal is to achieve an F.O.M. improvement of 10x

The initial charge is: to build **a system that is 10x more capable than Frontera**

PERFORMANCE OF AN APPLICATION

- We have essentially four factors in Application Performance:
  - Did the runtime change? (An analog to Strong Scaling – run the same problem in less time).
  - Did the problem size change? (An analog to Weak Scaling – run larger problems in fixed time)
  - Did we use more or less of the total resource? (An analog to Throughput).
  - Did the Physics change? (No good analog).
- Note we aren't *exactly* applying the scaling concepts from "traditional" benchmarking – a strong scaling plot by definition looks at changes in node counts on a single homogeneous system, but the notion applies.

PERFORMANCE OF AN APP

- We define $\Delta perf_i$, therefore, to be the product of four factors:
  - ΔT – The Change in Runtime from Frontera to the new System.
  - ΔS – The Change in problem size from Frontera to the new System
  - ΔE – (Ensemble) The Change in the fraction of Frontera to the fraction of the new system used to achieve the benchmark.
  - ΔP – The Change in physics in an enhanced model (what fraction of operations per datum is added).
  - $\Delta perf_i = \Delta T \times \Delta S \times \Delta E \times \Delta P$ --- Average of $\boldsymbol{\Delta perf_i}$ **is our Ten-X**

CSA PROJECTS

General Area of Science	Application	Project Name
Astronomy and Astrophysics	Athena++	Astrophysical Fluid Dynamics at Exascale
	ChaNGa	Evolution of baryons and galaxies across the age of the Universe
	IceTray	Multi-Messenger Astrophysics with ICECUBE
	Enzo-E	Accelerating cosmological simulations of the first galaxies through deep learning
Biophysics and Biology	NAMD	Molecular Mechanisms of Viral Infection
	ensemble	Viruses in Respiratory Aerosols
	WESTPA/AMBER	Phase Separation of Disordered Proteins
Computational Fluid Dynamics	ensemble	Multiphysics simulation of a full hypersonic vehicle
	PSDNS	Large-scale DNS
Geodynamics and Earth Systems	AWP-ODC	Seismic simulation for hazard management
	CESM2	Coupled Earth-Atmosphere models
	CM1	Supercell thunderstorms and tornado prediction
	ISSM	Data-driven and physics modeling of ice dynamics
	SeisSol	Off-fault inelastic processes and fluid effects in earthquake simulation
	rhea	Bridging short and long time scales in global plate tectonics
Materials Engineering	EPW	Quantum materials engineering at the exascale
	MuST	Electron localization in materials
	PARSEC	Quantum calculations for the optical and dielectric properties of aqueous liquids
Other Applications	Grover	Detecting misinformation and social biases
	MILC	Lattice QCD for flavor physics

CSA

- We strive to have a significant number of applications ready for Horizon on day 1
  - Performant on CPUs or GPUs
  - Performant at the scale of Horizon
  - Ready to hit Ten-X
- About 8 applications are GPU ready today (that includes some ML apps)
- Maybe 12 applications can be made GPU ready in a short(er) time span
- There will be applications that will not be ported to GPU
- Only a very small number of applications uses 'fancy' C++ today (Kokkos, SYCL, etc.)
- There are applications that have a complicated workflow
- Science = 'Large application' and many small ones; can all be ported?

HORIZON

- We are optimistic that we can win this
- Construction phase will start in the near future, so we hope
- We will be ready one day 1 at scale: CSA projects
- We will have a number of happy users spanning a wide range of science fields
- We are committed to making Horizon a success

Stay tuned


FRONTERA
TACC
NSF
TEXAS
lars@tacc.utexas.edu
6/3/23
31z